An overview of



Paul Burkhardt

hat is Big Data?

Readers have probably heard or read about Big Data, but what is it exactly? According to O'Reilly Media, the term was coined in 2005 and refers to "a wide range of large data sets almost impossible to manage and process using traditional management tools—due to their size, but also their complexity" [1]. Gartner defines it as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [2]. The International Data Corporation (IDC) states "Big data technologies describe a new generation of technologies and architectures, designed to economically extract <u>value</u> from very large <u>volumes</u> of a wide <u>variety</u> of data, by enabling high-velocity capture, discovery, and/or analysis" [3].

The three Vs—volume, velocity, and variety—are what many use to describe characteristics of Big Data. The consensus is that Big Data comes from many sources, is of any type, and the scale in both size and speed make it difficult to process and analyze. The National Institute of Standards and Technology (NIST) does not have an official definition of Big Data, but at their Joint Cloud and Big Data Workshop held in 2013, the NIST Director, Patrick Gallagher, agreed that key aspects of Big Data include notions of volume, velocity, and complexity. Gallagher went on to say, "We are really looking at a new paradigm, a place of data primacy where everything starts with consideration of the data rather than consideration of the technology" [4].

The Big Data era

If the last decade of computing could be envisioned as a journey through a landscape of rolling hills, dark forests, and winding paths, then a wayward traveler will encounter an uncertain and rough terrain, strewn with artifacts of IT efficiency and computer modernization, such as service-oriented architecture, grid computing, and web services. Below the surface are the fossils from the Big Iron era, huge mainframes that once bellowed and lumbered across the lands. At the end, the traveler will reach the Mountain of Data. It is steep, imposing, and so big that clouds shroud its peak, but yonder lies the horizon . . .

The beginning of the Big Data era is not marked by a definitive epoch, but it has been a persistent tide since the dot-com era which gave us the Amazon Elastic Compute Cloud (EC2), Wikipedia, Skype, and the mapping of the human genome. Google Trends indicate "big data" web searches were steady between 2004 and 2010 before it began increasing sharply. The September 30, 2012 issue of Research Trends found that the number of research publications on Big Data exceeded an

exponential rate of growth starting around 2008 (see figure 1).

As the web became a more important platform for both social and business needs, the amount of information began to grow dramatically. There were also significant advances in science which drove data production. The National Human Genome Research Institute reports that, in early 2008, the field of genomics developed second-generation sequencing platforms that began reducing the cost of DNA sequencing at a rate faster than Moore's Law—CPU performance doubles every 18 months—dramatically increasing the production of genomic data (see figure 2). The Large Hadron Collider spun up in September 2008, generating about 15 petabytes of data per year, leading to the 2012 discovery of a Higgs Boson particle.

The Big Data era is far from over. Smaller, smarter, web-enabled devices are becoming prevalent, from mobile phones to pacemakers. The first wireless pacemaker made by St. Jude Medical Inc. came online in 2011. Mundane appliances and systems such as your kitchen refrigerator and home security systems are also transmitting over the web. These interconnected devices and sensors make up the Internet of Things, resulting in greater information creation and consumption. The world is becoming more connected and mobile. The International Telecommunication Union predicts the number of

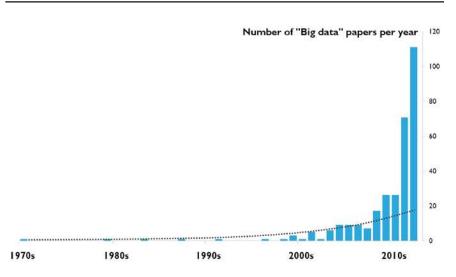


FIGURE 1. Research Trends found that the number of research publications on Big Data exceeded an exponential rate of growth starting around 2008. (Image from [5].)

mobile phone subscriptions will exceed the world population in 2014, which is over seven billion [7], and on January 2014, the IDC Worldwide Quarterly Mobile Phone Tracker reports that over one billion smartphones were shipped in a single year for the first time [8].

By 2017 the annual global Internet protocol traffic will top 1.4 zettabytes (ZB, i.e., 10²¹) [9], and by 2018 global mobile data traffic will reach 15.9 exabytes (EB, i.e., 1018) per month, according to Cisco (see figure 3) [10]. In 2014, Cisco also forecasts that 90% of all global mobile data traffic will be due to cloud applications by 2018 and that mobile cloud traffic will grow at a compound annual rate of 64% from 2013 to 2018 [10]. The forecast also predicts the number of mobile devices will reach 10 billion by 2018, with eight billion being personal devices, and over half of these mobile devices will

be smart devices accounting for more than 95% of global mobile data traffic (see figure 4) [10].

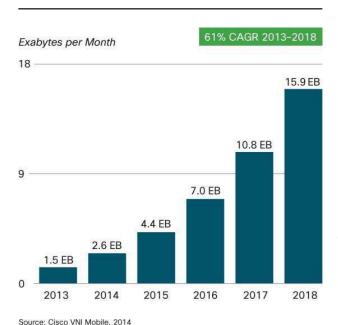


FIGURE 3. Cisco predicts that by 2018 global mobile data traffic will reach 15.9 EB per month, and the annual global Internet protocol traffic will top 1.4 ZB. (Image from "Cisco Visual Networking Index: Global mobile data traffic forecast update, 2013–2018" [10].)

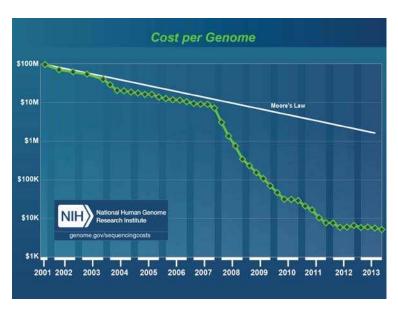
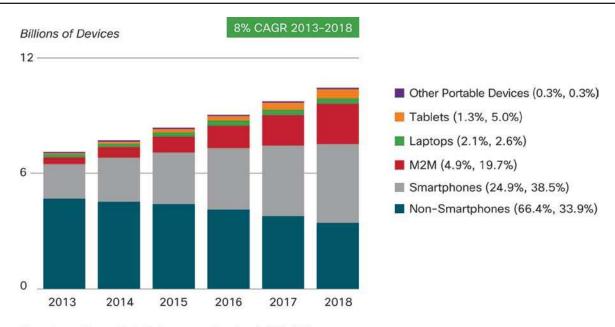


FIGURE 2. The National Human Genome Research Institute reports that, in early 2008, the field of genomics developed second-generation sequencing platforms that began reducing the cost of DNA sequencing at a rate faster than Moore's Law, dramatically increasing the production of genomic data. (Image from [6].)

How "big" is Big Data?

The growth of sensors and devices, coupled with social media and scientific breakthroughs, all sharing and transmitting over the same networks, contribute to the *data deluge*. According to an IDC study, "Extracting value from chaos," information is doubling every two years while metadata, data about data, is growing two times faster than data. A 2011 study of the world information capacity [11] estimated there were 295 EB of storage, 1.9 ZB of broadcast data (i.e., TV, radio) and 65 EB of telecommunication data (i.e., fixed phone, mobile phone, Internet)—in 2007.

The 2012 acting director of the Defense Advanced Research Projects Agency (DARPA), Kaigham J. Gabriel, used this analogy, "The Atlantic Ocean is roughly 350 million cubic kilometers in volume, or nearly 100 billion, billion gallons of water. If each gallon of water represented a byte or character, the Atlantic Ocean would be able to store, just barely, all the data generated by the world in 2010. Looking for a specific message or page in a document would be the equivalent of searching the Atlantic Ocean for a single 55-gallon drum barrel" [12].



Figures in parentheses refer to device or connections share in 2013, 2018. Source: Cisco VNI Mobile, 2014

FIGURE 4. Cisco predicts the number of mobile devices will reach 10 billion by 2018, with eight billion being personal devices, and over half of these mobile devices will be smart devices accounting for more than 95% of global mobile data traffic. (Image from "Cisco Visual Networking Index: Global mobile data traffic forecast update, 2013–2018" [10].)

The total digital information capacity, according to a 2012 IDC study, will reach 40 ZB by 2020 [13]. This projects approximately 10^{22} bytes of digital information in less than 10 years. By comparison, the number of stars in the universe is on order of 10^{22} , and the number of atoms in a mole is on order of 10^{23} (i.e., Avogadro's number). We are living in an era in which we have more data than resources to store and analyze it all.

Big Data in government

In March 2012, the White House announced the National Big Data Research and Development Initiative [14] to help address challenges facing the government, in response to the President's Council of Advisors on Science and Technology, which concluded the "Federal Government is underinvesting in technologies related to Big Data." With a budget of over \$200 million and support of six federal departments and agencies, this initiative was created to:

 Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data;

- ▶ Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

As part of the Big Data Initiative, the National Science Foundation (NSF) and the National Institutes of Health are funding a joint Big Data solicitation to "advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large and diverse data sets." In addition, the NSF is funding the \$10 million Expeditions in Computing project led by University of California at Berkeley, to turn data into knowledge and insight, and funding a \$2 million award for a research training group to support training for students in techniques for analyzing and visualizing complex data.

The Department of Defense (DoD) is also investing \$250 million annually to "harness and utilize massive data in new ways" and another \$60 million for new research proposals. DARPA, the research arm of the DoD, will invest \$25 million annually

2

under its XDATA program for techniques and tools to analyze large volumes of data, including

- Developing scalable algorithms for processing imperfect data in distributed data stores, and
- Creating effective human-computer interaction tools for facilitating rapidly customizable visual reasoning for diverse missions.

The Department of Energy is similarly providing \$25 million in funding to establish the Scalable Data Management, Analysis and Visualization Institute to develop new tools for managing and visualizing data.

What is all of this data?

Much of the "big" in Big Data comes unsurprisingly from web data. Web browsing activities are tracked in clickstream data that can be used to market the latest fads and trends. Social media also populates web data where Facebook, Twitter, and YouTube enable users to easily post, upload, and tweet about their daily activities.

Over 90% of digital information is unstructured [5]—data which has not been converted for organized analysis. Unstructured data is raw information created by ad hoc activities and can consist of images, video, logs, and documents. Unstructured data can be rich in semantic and relational information that must be parsed, interpreted, and transformed into canonicalized formats to fit neatly into catalogs and databases. In 2012, IDC believed 23% of information in our "digital universe" has Big Data value, increasing to one third by 2020, but only if the data is tagged and analyzed.

Who is creating all of this data?

Individuals are responsible for 75% of all digital information [15], half of which is generated *by* an individual through their activities, such as e-mails, phone calls, vacation photos, and video uploads. The other half is generated *about* an individual, a person's digital shadow, such as their web-browsing habits, financial transactions, surveillance footage, medical database, and GPS tracking.

Big Data economy

Big Data is Big Business—a January market report by Transparency Market Research claims the 2012 global Big Data market was worth \$6.3 billion and will grow to \$48.3 billion by 2018 [16], while IDC forecasts Big Data market revenue will grow at 31.7% per year, reaching \$23.8 billion in 2016 [17]. This global economy includes cloud technologies centered around Apache Hadoop software, which was inspired by Google's approaches to solving their internal Big Data dilemma. The value inherent in data can only be unlocked through analysis, which requires the right tools and infrastructure, but traditional methods are inadequate for Big Data, creating the necessity for innovation.

Can we process Big Data?

The volume and velocity of Big Data is exceeding our rate of physical storage and computing capacity, creating scalability demands that far outpace hardware innovations. Just as multicore chips were designed in response to the limits of clock speeds imposed by Moore's Law, cloud technologies have surfaced to address the impending tidal wave of information. The new cloud architectures pioneered by Google and Amazon extended distributed computing from its roots in high-performance computing and grid computing, where hardware was expensive and purpose-built, to large clusters made from low-cost commodity computers, ushering the paradigm of "warehouse" computing. These new cloud data centers containing thousands of computer cabinets are patrolled by administrators on motorized carts to pull and replace failed components.

On the software side, new parallel programming frameworks, like Hadoop MapReduce, help us crunch vast batches of data locally and independently by minimizing data dependency and the cost of transferring data between systems. Relational databases are being replaced by NoSQL and "eventually consistent" (key, value) stores such as Accumulo and MongoDB in order to scale with data demands.



Is Big Data secure?

The studies from IDC found that less than a third of the digital information was secured, and of the data requiring protection, only 20% was actually protected [3]. More than 80% of data about an individual will have passed through a commercial organization, a liability that is quietly ignored [3]. The lack of security is partly due to deference, but as data becomes more ubiquitous and available in commercial clouds, the need for security will become more imperative. Currently, the only open-source cloud database that offers cell-level security is Accumulo, a BigTable variant initially developed here at the National Security Agency (NSA).

Past the mountain of data

The DoD is leveraging the cloud infrastructure to consolidate 15,000 military networks, a potential annual savings of \$680 million [18]. At NSA, the Big Data challenge is compounded by the need for security and compliance where data must be compartmentalized and audited as part of the oversight requirements to protect the Fourth Amendment.

This requires cell-level security and provenance. The challenges for NSA are not limited to infrastructure; analysis is still the primary tradecraft of the Agency. The mountain of data must be processed into intelligence products that will safeguard national security. The need for Big Data analytics will be at the core of this tradecraft.

In today's Big Data era, information is exploding as networks converge and more devices come online in the Internet of Things. The scale of Big Data poses immense challenges, and with our unrelenting consumption of information, this is likely to persist. Addressing Big Data challenges will require efforts from government, academia, and industry.

About the author

Paul Burkhardt is a computer science researcher in the Research Directorate at NSA. He received his PhD from the University of Illinois at Urbana-Champaign. His current research interests are primarily focused on graph algorithms and Big Data analytics.

References

- [1] O'Reilly. "Roger Magoulas on Big Data" [YouTube video]. Uploaded 2010 Nov 16. Available at: http://youtu.be/fsT2NyM5BTI.
- [2] Beyer M, Laney D. "The importance of 'Big Data': A definition." Gartner. 2012 Jun 21.
- [3] Gantz J, Reinsel D. "Extracting value from chaos." IDC. 2011 Jun.
- [4] Gallagher P. [Remarks from] NIST Joint Cloud and Big Data Workshop. Jan 2013, Gaithersburg, MD. Webcast available at: http://www.nist.gov/itl/cloud/nist-joint-cloud-and-big-data-workshop-webcast.cfm.
- [5] Research Trends. "The evolution of Big Data as a research and scientific Topic: Special issue on Big Data." 2012 Sep 30. Available at: http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf.
- [6] Wetterstrand KA. "DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP)" [accessed 2014 Mar 31]. Available at: https://www.genome.gov/sequencingcosts.

- [7] International Telecommunication Union. "The world in 2013: ICT facts and figures." 2013 Feb. Available at: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf.
- [8] International Data Corporation. "Worldwide smartphone shipments top one billion units for the first time, according to IDC" [Press release]. 2014 Jan 27. Available at: http://www.idc.com/getdoc.jsp?containerId=prUS24645514.
- [9] Cisco. "The zettabyte era—trends and analysis." 2013 May 29. Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html.
- [10] Cisco. "Cisco Visual Networking Index: Global mobile data traffic forecast update, 2013–2018." 2014 Feb 5. Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf.

- [11] Hilbert M, Lopez P. "The world's technology capacity to store, communicate and compute information." *Science*. 2011;332(6025):60–65. doi: 10.1126/science.1200970.
- [12] Defense Advanced Research Projects Agency. "DARPA calls for advances in 'Big Data' to help the warfighter" [Press release]. 2012 Mar 29. Available at: http://www.darpa.mil/NewsEvents/Releases/2012/03/29.aspx.
- [13] Gantz J, Reinsel D. "The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the Far East." International Data Corporation. 2012 Dec. Available at: http://idcdocserv.com/1414.
- [14] Office of Science and Technology Policy, Executive Office of the President. "Obama administration unveils 'Big Data' initiative: Announces \$200 million in new R&D investments" [Press release]. 2012 Mar 29. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

- [15] Gantz JF, Chute C, Manfrediz A, Minton S, Reinsel D, Schlichting W, Toncheva A. "The diverse and exploding digital universe." 2008 Mar. International Data Corporation. Available at: http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.
- [16] Transparency Market Research. "Big Data market—Global scenario, trends, industry analysis, size, share and forecast 2012–2018." 2013 Jan 21.
- [17] International Data Corporation. "New IDC Big Data technology and services forecast shows worldwide market expected to reach to \$23.8 billion in 2016" [Press release]. 2013 Jan 08. Available at: http://www.idc.com/getdoc.jsp?containerid=prUS23900013.
- [18] Sternstein A. "NSA chief endorses the cloud for classified military cyber program." 2012 Jun 13. *Nextgov.* Available at: http://www.nextgov.com/ cybersecurity/2012/06/nsa-chief-endorses-cloud-classifiedmilitary-cyber-program/56257/.

The Next Wave is published to disseminate technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the US Government.

This publication is available online at http://www.nsa.gov/research/tnw/index.shtml. For more information, please contact us at TNW@tycho.ncsc.mil.